**1Д**

**УДК 004.912**

***Г.В. Дорохина*[1], *В.Ю. Трунов*[2], *Е.В. Шилова*[2]**
[1] Институт проблем искусственного интеллекта МОН Украины и НАН Украины,
г. Донецк, Украина
[2] Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
sgv@iai.donetsk.ua

# Модуль морфологического анализа без словаря слов русского языка

Статья посвящена созданию программного обеспечения бессловарного морфологического анализа слов русского языка. В работе описан разработанный программный модуль, характеристики его функционирования, область и перспективы применения.

## Введение

Морфологический анализ (МА) позволяет определить принадлежность словоформы к определенной лексеме (поиск леммы) и грамматические признаки словоформы – морфологическую информацию (МИ). МА является неотъемлемой частью систем анализа естественно-языковых текстов и интеллектуальных информационно-поисковых систем.

Системы морфологического анализа (МА) со словарем обеспечивают «более полный анализ словоформы», но «на пространстве реальных текстов» они «часто дают сбои», так как «не существует полных словарей» [1]. Для компенсации этого недостатка для слов, отсутствующих в словаре системы МА, целесообразно использование МА без словаря (БМА). Результаты БМА также можно использовать для неточной, но быстрой оценки морфологической информации словоформы при поверхностном синтаксическом анализе.

В ходе предшествующих работ ИПИИ в данном направлении был создан модуль декларативного морфологического анализа (ДМА) слов русского языка [2]. Он в явном виде хранит парадигмы слов (около 3 млн словоформ, из которых 1,9 млн уникальных строк) и представляет собой совокупность средств морфологического анализа и синтеза. Каждая словоформа снабжена МИ. Модуль использует метод скоростного поиска строковых величин в словарях сверхбольших объёмов [3]. Нерешенными проблемами этой работы является обработка слов, не представленных в словаре, и отсутствие подсистемы автоматизации пополнения словаря. Для их решения необходимо создание системы БМА с функциональными возможностями морфологического анализа и синтеза, что обосновывает **актуальность** работы.

**Цель работы** – разработка программного обеспечения бессловарного морфологического анализа слов русского языка со следующими функциональными возможностями: нахождение леммы и морфологической информации, нахождение словоформы или парадигмы для введённой строки.

# Описание программного комплекса

Модуль БМА имеет 2 режима функционирования: режим наполнения (РН) и режим функционирования (РФ).

Для наполнения модуля используем словарную базу модуля ДМА. Это позволяет учесть при МА без словаря правила словоизменения всех слов, включенных в данный модуль. В режиме наполнения на вход модуля БМА последовательно подаются парадигмы, представленные массивом строк (словоформы) и чисел (МИ). При этом выходных данных модуль БМА не имеет.

В качестве базы данных модуля БМА используем базу **парадигм псевдоокончаний**. Для пояснения данного термина необходимо определить несколько понятий.

**Парадигма** – множество грамматических форм слова, каждая из которых описывается с помощью МИ, а также множество написаний слова в этих формах (**словоформ**).

**Псевдооснова** – максимально длинная общая начальная часть словоформ парадигмы.

**Псевдоокончание** – это последовательность символов словоформы, из которой удалена псевдооснова парадигмы кроме последней буквы псевдоосновы. Тогда **парадигма псевдоокончаний** – совокупность псевдоокончаний и значений МИ всех грамматических форм слова.

В силу специфики МА без словаря, в режиме функционирования многие результаты «на выходе» системы могут не являться словами русского языка. Поэтому выходные данные системы необходимо ранжировать (присваивать более вероятным результатам больший ранг) и выводить в порядке убывания ранга.

Таким образом, модуль БМА имеет 4 информационных и 1 управляющий вход, а также 4 информационных выхода (рис. 1).

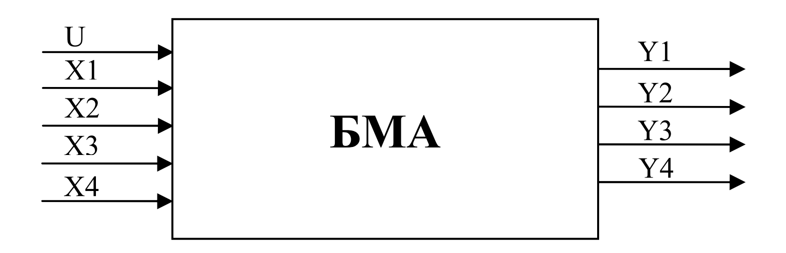



Рисунок 1 – Схема функционирования модуля БМА

Режим работы и функции модуля БМА в зависимости от управляющего сигнала, используемые при этом входные и генерируемые выходные сигналы приведены в табл. 1.

В общем виде функционирование системы можно описать следующим образом. Модуль БМА использует хранилище строк [3], и таблицу парадигм псевдоокончаний. Хранилище строк содержит инвертированные написания псевдоокончаний и позволяет определить идентификаторы псевдоокончаний, которые потенциально присущи введенной строке. Найденные идентификаторы позволяют определить номера соответствующих записей таблицы парадигм. А они, в свою очередь, используются для нахождения МИ, потенциальной леммы и парадигмы слова.

Таблица 1 – Соотношение входных, выходных данных и функций модуля

<table>
<tr><td rowspan="3">Режим</td><td colspan="4">Входные данные БМА</td><td colspan="2" rowspan="2">Выходные данные БМА</td></tr>
<tr><td colspan="2">Управляющий сигнал (U)</td><td colspan="2">Данные</td></tr>
<tr><td>значение</td><td>описание</td><td>вход</td><td>тип данных (описание)</td><td>выход</td><td>тип данных (описание)</td></tr>
<tr><td rowspan="3">РФ</td><td>1</td><td>Найти лемму</td><td>X1</td><td>строка (словоформа)</td><td>Y1<br>Y2<br>Y3<br>Y4</td><td>массив строк (леммы)<br>массив чисел (МИ)<br>массив чисел (№ парадигмы)<br>массив чисел (ранг)</td></tr>
<tr><td>2</td><td>Найти парадигму</td><td>X1</td><td>строка (лемма)</td><td>Y1<br>Y2<br>Y3<br>Y4</td><td>массив строк (словоформ)<br>массив чисел (МИ)<br>массив чисел (№ парадигмы)<br>массив чисел (ранг)</td></tr>
<tr><td>3</td><td>Найти словоформу</td><td>X1<br>X2</td><td>строка (лемма)<br>число (МИ)</td><td>Y1<br>Y2<br>Y3<br>Y4</td><td>массив строк (словоформ)<br>массив чисел (МИ)<br>массив чисел (№ парадигмы)<br>массив чисел (ранг)</td></tr>
<tr><td>РН</td><td>4</td><td>Добавить парадигму</td><td>X3<br>X4</td><td>массив строк (словоформы)<br>массив чисел (МИ)</td><td></td><td></td></tr>
</table>

Важным этапом разработки модуля БМА является наполнение словарной базы. Наполнение словарной базы состоит в:

– последовательном выборе парадигм из словарной базы модуля ДМА;

– построении для каждой из них парадигмы псевдоокончаний;

– поиске в словарной базе полученной парадигмы псевдоокончаний;

– добавлении отсутствующих парадигм псевдоокончаний в базу данных или наращивании ранга найденной парадигмы.

Данные операции выполняем для парадигм изменяемых слов из словарной базы модуля ДМА, написание которых не содержит дефиза и строка псевдоосновы которых имеет ненулевую длину.

Поиск в базе БМА некоторой парадигмы псевдоокончаний состоит в последовательном выборе парадигм псевдоокончаний и сравнении их с искомой парадигмой. При этом парадигмы псевдоокончаний считаются одинаковыми, если у них совпадает перечень грамматических форм, а формы с одинаковой МИ имеют одинаковые псевдоокончания. Для ускорения процедуры сравнения парадигм псевдоконачаний они должны быть упорядочены по возрастанию значений МИ.

Время, необходимое для проверки наличия парадигмы псевдоокончаний в базе модуля БМА, тем больше, чем больше парадигм она содержит. Поэтому формирование базы данных модуля БМА выполняем в 2 этапа: 1) создаем базы парадигм псевдоокончаний для отдельных частей речи; 2) объединяем базы парадигм псевдоокончаний отдельных частей речи.

Нахождение леммы можно описать так. Получаем массив идентификаторов псевдоокончаний. Для каждого из них находим массив номеров записей таблицы парадигм. Для каждой из этих записей формируем строку леммы и запоминаем морфологическую информацию.

Процесс нахождения парадигмы аналогичен. Парадигмы находятся только для тех записей таблицы парадигм, которые являются леммами. Для них находим парадигмы псевдоокончаний, по которым формируем массив словоформ и МИ.

## Анализ результатов функционирования модуля БМА

Поскольку модуль БМА предназначен для использования совместно с модулем ДМА, нет необходимости вносить в него служебные части речи, местоимения и числительные, наречия.

База данных модуля БМА заполнена парадигмами псевдоокончаний существительных, прилагательных и глаголов. Результирующая база включает в себя около 19 тыс. парадигм псевдоокончаний (таблица парадигм из 680 тыс. записей, около 32,5 тыс. различных псевдоокончаний).

Использование модуля БМА в задаче морфологического парсинга текста[1] позволило оценить среднее время МА модулем. На множестве 166 260 слов среднее время анализа строки составило 0,013533 с. Для сравнения отметим, что модуль ДМА обрабатывает строку за 0,000378 с (на множестве 3 575 666 слов). Оценка среднего времени анализа строки проводилась в рамках одного выполнения программы морфологического парсинга, причём обращение к модулю БМА производилось в случае неудачи поиска леммы модулем ДМА (то есть выполнялось попеременное обращение к модулям ДМА и БМА).

Более чем 35-кратное увеличение времени анализа строки модулем БМА, очевидно, связано с получаемым в процессе работы модуля БМА большим количеством повторов результатов МА (совпадают МИ и леммы) и необходимостью их удаления из множества выходных данных. В дальнейшем целесообразно направить усилия на ускорение МА модулем БМА.

Большое количество повторов результатов МА, получаемых в процессе работы модуля, указывает на принадлежность псевдоокончания с заданным написанием большому количеству парадигм псевдоокончаний. Поскольку вошедшие в базу парадигмы отличаются друг от друга количеством, перечнем, написанием псевдоокончаний (одного или нескольких), при нахождении парадигмы необходимо синтезировать парадигмы входного слова с учётом каждой из них. Таким образом, «на выходе» системы получаем чрезмерно большое количество парадигм, которое применительно к задаче пополнения словаря ставит проблему выбора добавляемой в словарь парадигмы.

## Выводы

В результате работы создан модуль БМА слов русского языка, который применим к обработке слов, не представленных в словаре. Так, при описанном в работе анализе корпуса текстов необходимость в его применении возникла в 4,5% случаев.

[1] В рамках Форума «Оценка методов автоматического анализа текста: морфологические парсеры русского текста» (Режим доступа : http://ru-eval.ru/).

Значительный рост временных затрат модуля БМА на анализ словоформы в сравнении с модулем ДМА указывает на необходимость его совершенствования в этом направлении.

Относительно использования модуля БМА в качестве средства автоматизации пополнения словаря приходим к выводу о проблематичности пополнения словаря ДМА путём построения потенциальных парадигм одной словоформы, в силу их большого их количества. Более эффективным представляется анализ построенных модулем БМА множеств потенциальных парадигм «несловарных» слов некоторого текста или корпуса текстов и принятие решения о парадигмах этих слов. Эта задача может стать одним из направлений развития работы.

# Литература

1. Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы) [Электронный ресурс] / И.М. Ножов. – Москва, 2003. – Internet-публикация содержит исправления и сокращения оригинального текста диссертации «Реализация автоматической синтаксической сегментации русского предложения». – Режим доступа : www.aot.ru/docs/Nozhov/msot.pdf
2. Дорохина Г.В. Модуль морфологического анализа слов русского языка / Г.В. Дорохина, А.П. Павлюкова // Искусственный интеллект. – 2004. – № 3. – С. 636-642.
3. Пат. №78806 Україна, МПК G 06 F 17/30, G 06 F 7/76, G 06 F 12/00. Пристрій для збереження і пошуку рядкових величин та спосіб збереження і пошуку рядкових величин / Дорохіна Галина Володимирівна ; заявник і власник Інститут проблем штучного інтелекту. – № a200500327 ; заявл. 14.01.2005 ; опубл. 25.04.2007

***Дорохіна Г.В., Трунов В.Ю., Шилова Є.В.***

**Модуль морфологічного аналізу без словника слів російської мови**

Статтю присвячено створенню програмного забезпечення безсловникового морфологічного аналізу слів російської мови. У роботі описано розроблений програмний модуль, характеристики його функціонування, область та перспективи використання.

***G.V. Dorokhina, V.Yu. Trunov, E.V. Shilova***

**Russian Words Morphological Analysis without Dictionary Module**

The article is devoted to creation of software for Russian words morphological analysis without a dictionary. The developed program unit, the descriptions of it’s functioning, the area and prospects of it’s application are described.

*Статья поступила в редакцию 25.03.2010.*